MEM第二次作业

Xu Dan

2024-11-20

Question #1

BigBangTheory. (Attached Data: BigBangTheory)

The Big Bang Theory, a situation comedy featuring Johnny Galecki, Jim Parsons, and Kaley Cuoco-Sweeting, is one of the most-watched programs on network television. The first two episodes for the 2011–2012 season premiered on September 22, 2011; the first episode attracted 14.1 million viewers and the second episode attracted 14.7 million viewers. The attached data file BigBangTheory shows the number of viewers in millions for the first 21 episodes of the 2011–2012 season (*the Big Bang theory* website, April 17, 2012).

a. Compute the minimum and the maximum number of viewers.

The minimum number of viewers is 13.3.

The maximum number of viewers is 16.5.

```
min(data1$`Viewers (millions)`)
```

```
## [1] 13.3
```

```
max(data1$`Viewers (millions)`)
```

```
## [1] 16.5
```

b. Compute the mean, median, and mode.

The mean is 15.

The median is 15.

The mode is 13.6

```
mean(data1$`Viewers (millions)`)
```

```
## [1] 15.04286
```

```
median(data1$`Viewers (millions)`)
```

```
## [1] 15
```

```
names(table(data1$`Viewers (millions)`))[which.max(table(data1$`Viewers (millions)`))]
```

```
## [1] "13.6"
```

c. Compute the first and third quartiles.

The first quartile is 14.

The third quartile is 16.

```
quantile(data1$`Viewers (millions)`,0.25)
```

```
##  25%
## 14.1
```

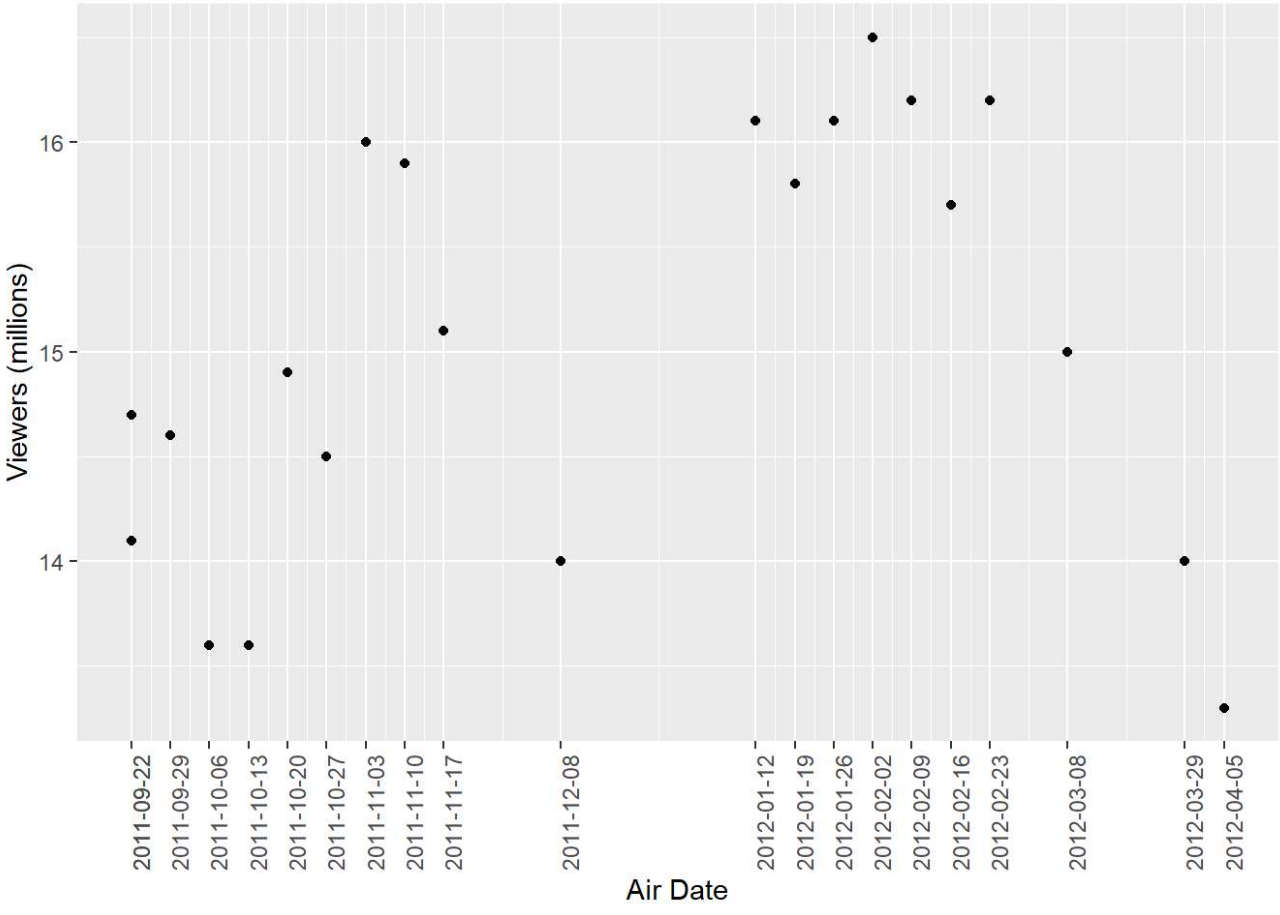
```
quantile(data1$`Viewers (millions)`,0.75)
```

```
## 75%
##  16
```

d. has viewership grown or declined over the 2011–2012 season? Discuss.

From the plot blow, there is no trend in 2011 season, while there is a booming at the beginning of the 2012 season and the number of viewers steep declined at the of the season.

```
ggplot(data1)+
  geom_point(aes(x=`Air Date`,y=`Viewers (millions)`))+
  scale_x_date(breaks = data1$`Air Date`) +
  theme(axis.text.x = element_text(angle = 90))
```



Question #2

NBAPlayerPts. (Attached Data: NBAPlayerPts)

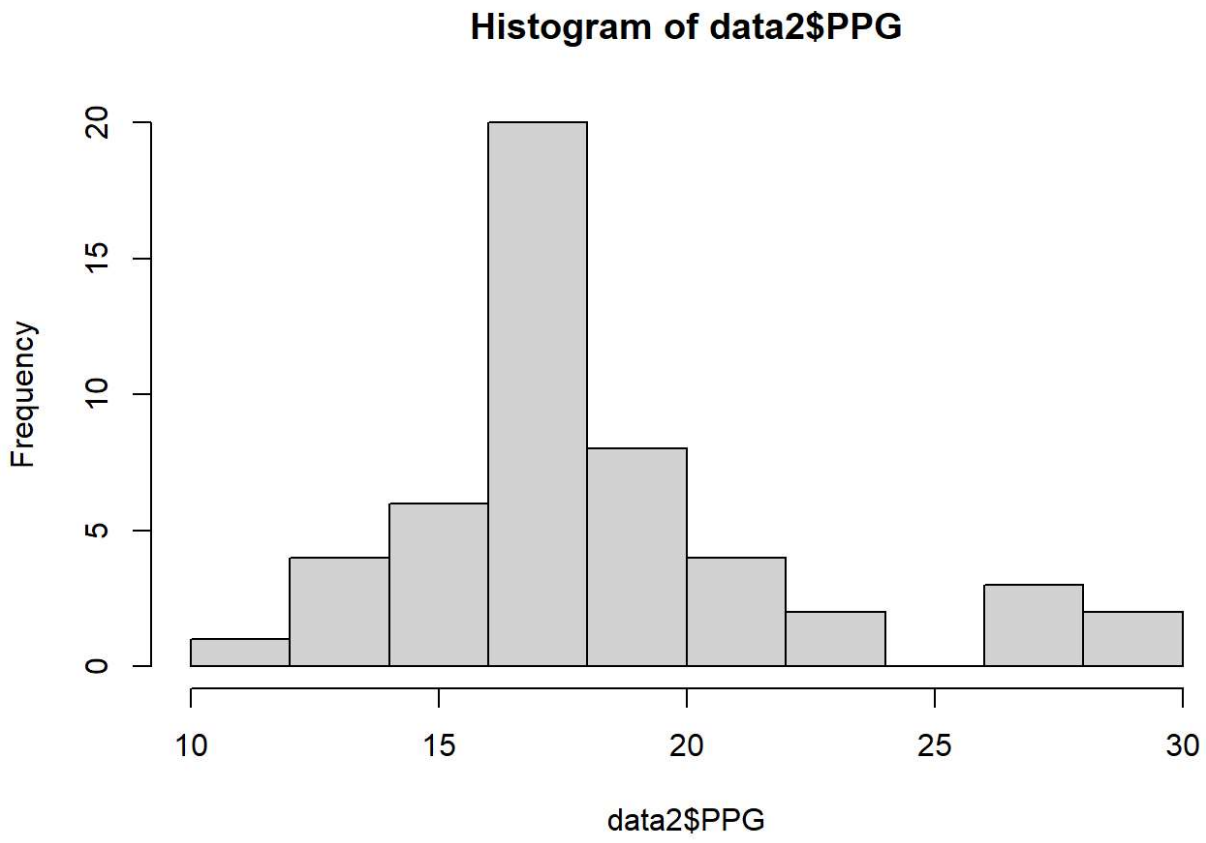
CbSSports.com developed the Total Player Rating system to rate players in the National Basketball Association (NBA) based on various offensive and defensive statistics. The attached data file NBAPlayerPts shows the average number of points scored per game (PPG) for 50 players with the highest ratings for a portion of the 2012–2013 NBA season (CbSSports.com website, February 25, 2013). Use classes starting at 10 and ending at 30 in increments of 2 for PPG in the following.Show the frequency distribution.

a.

```
table(cut_width(data2$PPG,2,boundary = 10))
```

```
##
## [10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
##       1       4       6      20       8       4       2       0       3       2
```

```
hist(data2$PPG,breaks = seq(10, 30, by = 2))
```

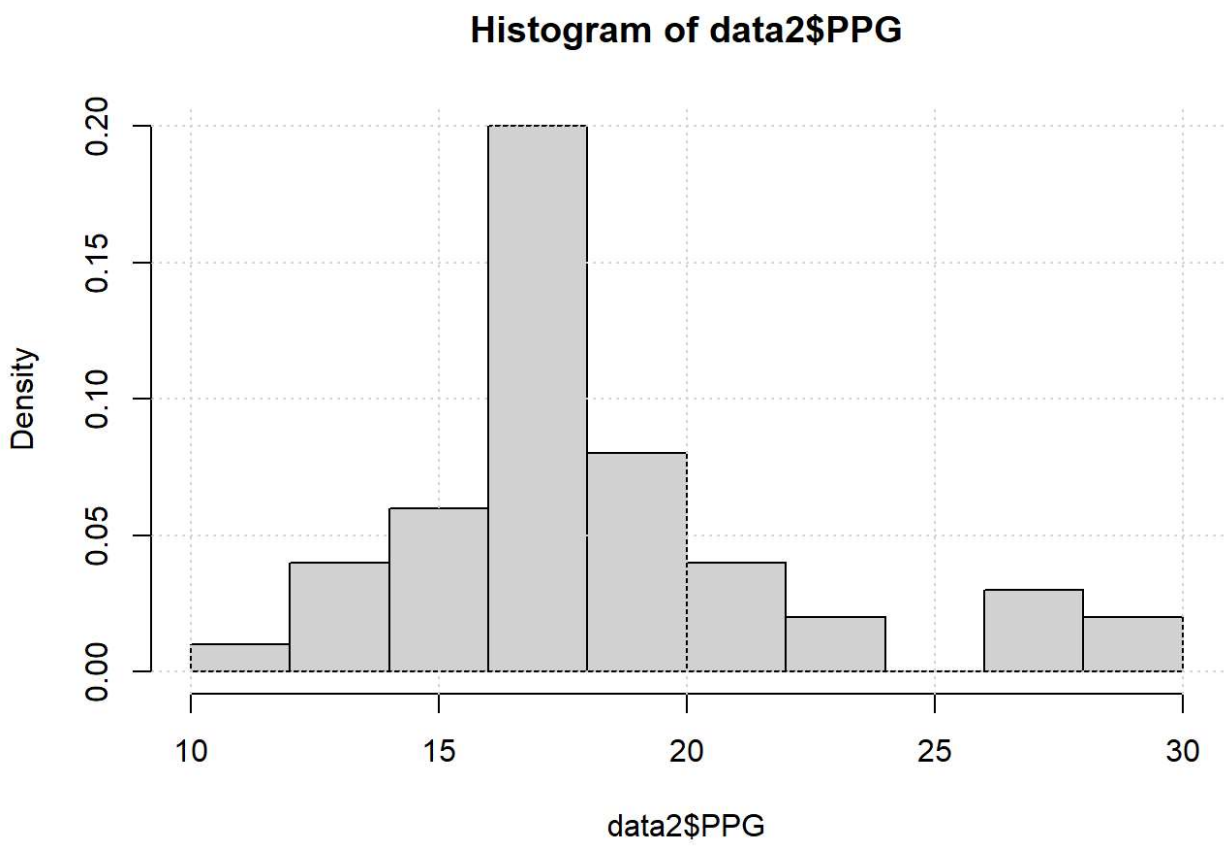


b. Show the relative frequency distribution.

```
table(cut_width(data2$PPG,2,boundary = 10))/50
```

```
##
## [10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
##    0.02    0.08    0.12    0.40    0.16    0.08    0.04    0.00    0.06    0.04
```

```
hist(data2$PPG,probability = TRUE,breaks = seq(10, 30, by = 2))
grid()
```

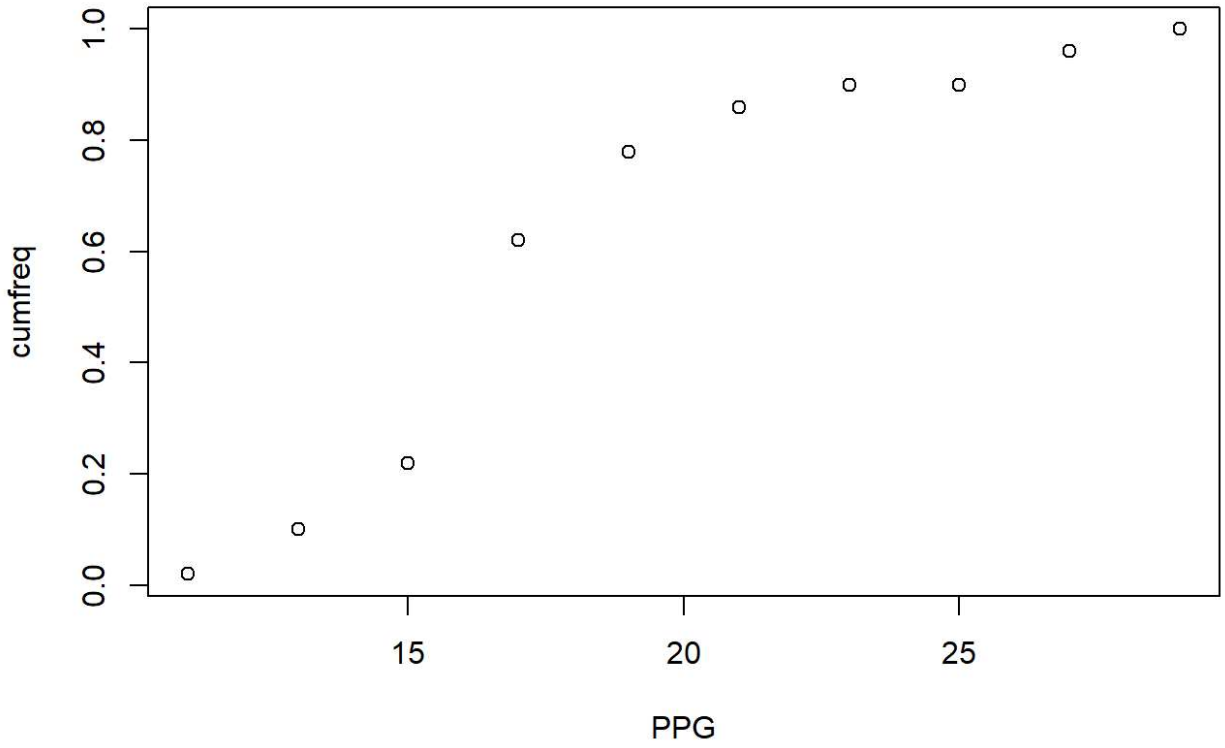


c. Show the cumulative percent frequency distribution.

```
cumsum(table(cut_width(data2$PPG,2,boundary = 10))/50)
```

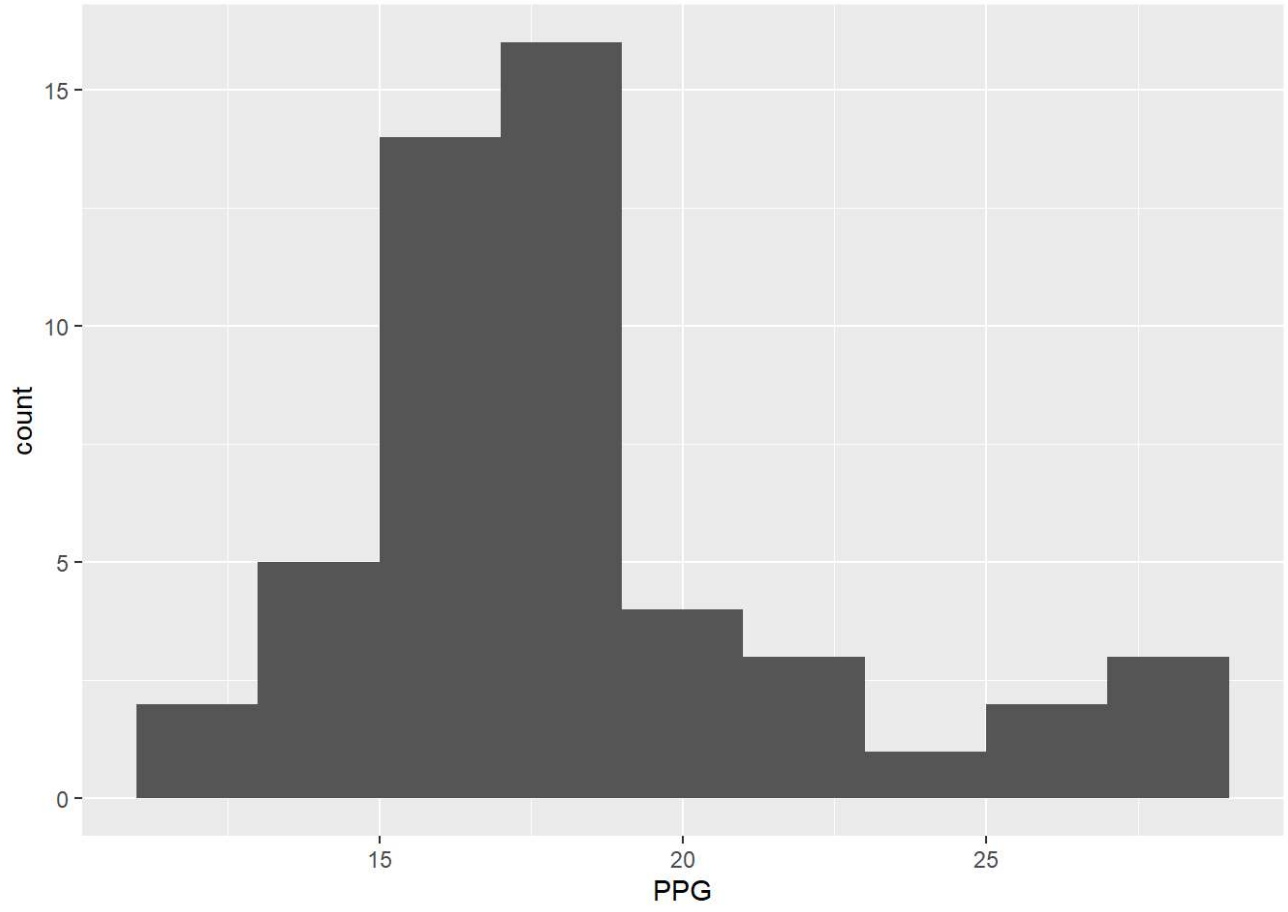
```
## [10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
##    0.02    0.10    0.22    0.62    0.78    0.86    0.90    0.90    0.96    1.00
```

```
hist_info <- hist(data2$PPG, breaks = seq(10, 30, by = 2), plot = FALSE)
cumfreq <- cumsum(hist_info$counts) / sum(hist_info$counts)
plot(hist_info$mids, cumfreq,xlab = "PPG")
```



d. Develop a histogram for the average number of points scored per game.

```
ggplot(data2,aes(PPG)) +
  geom_histogram(binwidth = 2)
```



e. Do the data appear to be skewed? Explain.

Skewed rightly.
From the plot above, it has a long tail to the right.

f. What percentage of the players averaged at least 20 points per game?

1-0.78=0.22

Question #3

A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.

a. How large was the sample used in this survey

$(500/20)^2 = 625$

b. What is the probability that the point estimate was within ±25 of the population mean?

The probability is 0.7887.

```
(pnorm(25/20)-0.5)*2
```

```
## [1] 0.7887005
```

Question #4

Young Professional Magazine (Attached Data: Professional)

Young Professional magazine was developed for a target audience of recent college graduates who are in their first 10 years in a business/professional career. In its two years of publication, the magazine has been fairly successful. Now the publisher is interested in expanding the magazine's advertising base. Potential advertisers continually ask about the demographics and interests of subscribers to *young Professionals*. To collect this information, the magazine commissioned a survey to develop a profile of its subscribers. The survey results will be used to help the magazine choose articles of interest and provide advertisers with a profile of subscribers. As a new employee of the magazine, you have been asked to help analyze the survey results.

Some of the survey questions follow:

1. What is your age?
2. Are you: Male_____ Female_____
3. Do you plan to make any real estate purchases in the next two years?

   Yes_____ No_____
4. What is the approximate total value of financial investments, exclusive of your

   home, owned by you or members of your household?
5. How many stock/bond/mutual fund transactions have you made in the past year?
6. Do you have broadband access to the Internet at home? Yes_____ No_____
7. Please indicate your total household income last year. _____
8. Do you have children? Yes_____ No_____

The file entitled Professional contains the responses to these questions.

Managerial Report:

Prepare a managerial report summarizing the results of the survey. In addition to statistical summaries, discuss how the magazine might use these results to attract advertisers. You might also comment on how the survey results could be used by the magazine's editors to identify topics that would be of interest to readers. Your report should address the following issues, but do not limit your analysis to just these areas.

a. Develop appropriate descriptive statistics to summarize the data.

```
library(psych)
```

```
##
## 载入程序包：'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
describe(data4)
```

```
## Warning in FUN(newX[, i], ...): min里所有的参数都不存在；返回Inf
## Warning in FUN(newX[, i], ...): min里所有的参数都不存在；返回Inf
## Warning in FUN(newX[, i], ...): min里所有的参数都不存在；返回Inf
## Warning in FUN(newX[, i], ...): min里所有的参数都不存在；返回Inf
## Warning in FUN(newX[, i], ...): min里所有的参数都不存在；返回Inf
```

```
## Warning in FUN(newX[, i], ...): max里所有的参数都不存在；返回-Inf
## Warning in FUN(newX[, i], ...): max里所有的参数都不存在；返回-Inf
## Warning in FUN(newX[, i], ...): max里所有的参数都不存在；返回-Inf
## Warning in FUN(newX[, i], ...): max里所有的参数都不存在；返回-Inf
## Warning in FUN(newX[, i], ...): max里所有的参数都不存在；返回-Inf
```

```
##                           vars   n     mean       sd median  trimmed      mad
## Age                          1 410    30.11     4.02     30    30.14     4.45
## Gender*                      2 410     1.56     0.50      2     1.57     0.00
## Real Estate Purchases?*      3 410     1.44     0.50      1     1.43     0.00
## Value of Investments ($)     4 410 28538.29 15810.83  24800 26561.59 11267.76
## Number of Transactions       5 410     5.97     3.10      6     5.66     2.97
## Broadband Access?*           6 410     1.62     0.48      2     1.66     0.00
## Household Income ($)         7 410 74459.51 34818.21  66050 69666.16 26019.63
## Have Children?*              8 410     1.53     0.50      2     1.54     0.00
## ...9                         9   0      NaN       NA     NA      NaN       NA
## ...10*                      10   1     1.00       NA      1     1.00     0.00
## ...11                       11   0      NaN       NA     NA      NaN       NA
## ...12                       12   0      NaN       NA     NA      NaN       NA
## ...13                       13   0      NaN       NA     NA      NaN       NA
## ...14                       14   0      NaN       NA     NA      NaN       NA
##                            min    max  range  skew kurtosis      se
## Age                         19     42     23 -0.03    -0.01    0.20
## Gender*                      1      2      1 -0.23    -1.95    0.02
## Real Estate Purchases?*      1      2      1  0.23    -1.95    0.02
## Value of Investments ($)     0 133400 133400  1.70     5.52  780.84
## Number of Transactions       0     21     21  1.21     2.39    0.15
## Broadband Access?*           1      2      1 -0.51    -1.74    0.02
## Household Income ($)     16200 322500 306300  1.99     7.34 1719.55
## Have Children?*              1      2      1 -0.14    -1.99    0.02
## ...9                       Inf   -Inf   -Inf    NA       NA      NA
## ...10*                       1      1      0    NA       NA      NA
## ...11                      Inf   -Inf   -Inf    NA       NA      NA
## ...12                      Inf   -Inf   -Inf    NA       NA      NA
## ...13                      Inf   -Inf   -Inf    NA       NA      NA
## ...14                      Inf   -Inf   -Inf    NA       NA      NA
```

b. Develop 95% confidence intervals for the mean age and household income of subscribers.

```
t.test(data4$Age,conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data4$Age
## t = 151.52, df = 409, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  29.72153 30.50286
## sample estimates:
## mean of x
##   30.1122
```

```
t.test(data4$`Household Income ($)`,conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  data4$`Household Income ($)`
## t = 43.302, df = 409, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  71079.26 77839.77
## sample estimates:
## mean of x
##  74459.51
```

c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

We can conclude with 95% confidence that the proportion of subscribers who have broadband access at home is between 0.5753 and 0.6711. And, we can conclude with 95% confidence that the proportion of subscribers who have children is between 0.4846 and 0.5831.

```
table(as.factor(data4$`Broadband Access?`))
```

```
##
##  No Yes
## 154 256
```

```
prop.test(256,410)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  256 out of 410, null probability 0.5
## X-squared = 24.88, df = 1, p-value = 6.1e-07
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5753252 0.6710862
## sample estimates:
##         p
## 0.6243902
```

```
table(as.factor(data4$`Have Children?`))
```

```
##
##  No Yes
## 191 219
```

```
prop.test(219,410)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  219 out of 410, null probability 0.5
## X-squared = 1.778, df = 1, p-value = 0.1824
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4845521 0.5830908
## sample estimates:
##         p
## 0.5341463
```

d. Would *Young Professional* be a good advertising outlet for online brokers? Justify your conclusion with statistical data.

   Yes. Young Professional would be a good advertising outlet for online brokers.

   As we can see in the answer of the previous question, we can conclude with 95% confidence that the proportion of subscribers who have broadband access at home is between 0.5753 and 0.6711. The proportion is large, and subscribers can easily access to online brokers.

   From the descriptive statistics of the data, the mean of the approximate total value of financial investments is $28538.29, which means the subscribers have idle funds for additional transactions. The average number of transactions stands at roughly 6 per year, although numerous subscribers engage in significantly more transactions.

e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?

   Yes. This magazine would be a good place to advertise for companies selling educational software and computer games for young children.

   As we can see in the answer of the third question, we can conclude with 95% confidence that the proportion of subscribers who have children is between 0.4846 and 0.5831. The average age of subscribers is 30.11 and there is a large proportion of subscribers have idle funds, which makes it reasonable for parents to buy educational software and computer games to their kids.

f. Comment on the types of articles you believe would be of interest to readers of *Young Professional*.

   People who have interests in real estate purchases and financial investments.

Question #5

Quality Associate, Inc. (Attached Data: Quality)

Quality associates, inc., a consulting firm, advises its clients about sampling and statistical procedures that can be used to control their manufacturing processes. in one particular application, a client gave Quality associates a sample of 800 observations taken during a time in which that client's process was operating satisfactorily. the sample standard deviation for these data was .21; hence, with so much data, the population standard deviation was assumed to be .21. Quality associates then suggested that random samples of size 30 be taken periodically to monitor the process on an ongoing basis. by analyzing the new samples, the client could quickly learn whether the process was operating satisfactorily. when the process was not operating satisfactorily, corrective action could be taken to eliminate the problem. the design specification indicated the mean for the process should be 12. the hypothesis test suggested by Quality associates follows.

$$H_0 : \mu = 12$$
$$H_1 : \mu \neq 12$$

Corrective action will be taken any time H_0 is rejected.

Data are available in the data set Quality.

Managerial Report

a. Conduct a hypothesis test for each sample at the .01 level of significance and determine what action, if any, should be taken. Provide the p-value for each test.

```
z.test(data5$`Sample 1`, mu=12, stdev=0.21, conf.level = 0.99)[3]
```

```
## $p.value
## [1] 0.2810083
```

```
z.test(data5$`Sample 2`, mu=12, stdev=0.21, conf.level = 0.99)[3]
```

```
## $p.value
## [1] 0.4546503
```

```
z.test(data5$`Sample 3`, mu=12, stdev=0.21, conf.level = 0.99)[3]
```

```
## $p.value
## [1] 0.003790318
```

```
z.test(data5$`Sample 4`, mu=12, stdev=0.21, conf.level = 0.99)[3]
```

```
## $p.value
## [1] 0.03389336
```

b. compute the standard deviation for each of the four samples. does the assumption of .21 for the population standard deviation appear reasonable?

Yes. The assumption of .21 for the population standard deviation appears reasonable.

```
sd(data5$`Sample 1`)
```

```
## [1] 0.220356
```

```
sd(data5$`Sample 2`)
```

```
## [1] 0.220356
```

```
sd(data5$`Sample 3`)
```

```
## [1] 0.2071706
```

```
sd(data5$`Sample 4`)
```

```
## [1] 0.206109
```

c. compute limits for the sample mean $\overline{x}$ around $\mu = 12$ such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. if $\overline{x}$ exceeds the upper limit or if $\overline{x}$ is below the lower limit, corrective action will be taken. these limits are referred to as upper and lower control limits for quality control purposes.

```
12+qnorm(0.01)*0.21/sqrt(30)
```

```
## [1] 11.91081
```

```
12-qnorm(0.01)*0.21/sqrt(30)
```

```
## [1] 12.08919
```

d. discuss the implications of changing the level of significance to a larger value. what mistake or error could increase if the level of significance is increased?

Type I error will increase.

```
12+qnorm(0.05)*0.21/sqrt(30)
```

```
## [1] 11.93694
```

```
12-qnorm(0.05)*0.21/sqrt(30)
```

```
## [1] 12.06306
```

Question #6

Vacation occupancy rates were expected to be up during March 2008 in Myrtle Beach, South Carolina (*the sun news,* February 29, 2008). Data in the file Occupancy (Attached file **Occupancy**) will allow you to replicate the findings presented in the newspaper. The data show units rented and not rented for a random sample of vacation properties during the first week of March 2007 and March 2008.

a. Estimate the proportion of units rented during the first week of March 2007 and the first week of March 2008.

```
table(as.factor(data6$`Unit Rented?`))
```

```
##
## March 2007         No        Yes
##          1        130         70
```

```
table(as.factor(data6$...2))
```

```
##
## March 2008         No        Yes
##          1         80         70
```

```
70/200 #March 2007
```

```
## [1] 0.35
```

```
70/150 #March 2008
```

```
## [1] 0.4666667
```

b. Provide a 95% confidence interval for the difference in proportions.

The 95% confidence interval for the difference in proportions are -0.2262 and -0.0072.

```
x <- c(70,70)
n <- c(200,150)
prop.test(x,n)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 4.3872, df = 1, p-value = 0.03621
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.226151510 -0.007181823
## sample estimates:
##    prop 1    prop 2
## 0.3500000 0.4666667
```

c. On the basis of your findings, does it appear March rental rates for 2008 will be up

from those a year earlier?

The p-value is 0.04, which is smaller than the significant level, so we can reject the equality hypothesis.

Question #7

Air Force Training Program (data file: Training)

An air force introductory course in electronics uses a personalized system of instruction whereby each student views a videotaped lecture and then is given a programmed instruc-tion text. the students work independently with the text until they have completed the training and passed a test. Of concern is the varying pace at which the students complete this portion of their training program. Some students are able to cover the programmed instruction text relatively quickly, whereas other students work much longer with the text and require additional time to complete the course. The fast students wait until the slow students complete the introductory course before the entire group proceeds together with other aspects of their training.

A proposed alternative system involves use of computer-assisted instruction. In this method, all students view the same videotaped lecture and then each is assigned to a computer terminal for further instruction. The computer guides the student, working independently, through the self-training portion of the course.

To compare the proposed and current methods of instruction, an entering class of 122 students was assigned randomly to one of the two methods. one group of 61 students used the current programmed-text method and the other group of 61 students used the proposed computer-assisted method. The time in hours was recorded for each student in the study. Data are provided in the data set training (see Attached file).

Managerial Report

a. use appropriate descriptive statistics to summarize the training time data for each method. what similarities or differences do you observe from the sample data?

```
library(psych)
describe(data7)
```

```
##          vars  n  mean   sd median trimmed  mad min max range  skew kurtosis
## Current     1 61 75.07 3.94     76   75.12 4.45  65  84    19 -0.21    -0.25
## Proposed    2 61 75.43 2.51     76   75.49 1.48  69  82    13 -0.27     0.33
##            se
## Current  0.51
## Proposed 0.32
```

b. Comment on any difference between the population means for the two methods. Discuss your findings.

There is no difference between the two groups in the significant level of 0.05.

```
t.test(data7$Current,data7$Proposed)
```

```
##
##  Welch Two Sample t-test
##
## data:  data7$Current and data7$Proposed
## t = -0.60268, df = 101.65, p-value = 0.5481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5476613  0.8263498
## sample estimates:
## mean of x mean of y
##  75.06557  75.42623
```

c. compute the standard deviation and variance for each training method. conduct a hypothesis test about the equality of population variances for the two training methods. Discuss your findings.

The p-value is much smaller than the the significant level of 0.05, we can reject the null hypothesis. The Current method has larger variance.

```
sd(data7$Current)
```

```
## [1] 3.944907
```

```
var(data7$Current)
```

```
## [1] 15.5623
```

```
sd(data7$Proposed)
```

```
## [1] 2.506385
```

```
var(data7$Proposed)
```

```
## [1] 6.281967
```

```
var.test(data7$Current,data7$Proposed)
```

```
##
##  F test to compare two variances
##
## data:  data7$Current and data7$Proposed
## F = 2.4773, num df = 60, denom df = 60, p-value = 0.000578
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.486267 4.129135
## sample estimates:
## ratio of variances
##           2.477296
```

d. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain.

   The proposed method is superior. The two methods have similar mean completion times. However, the proposed method has a significantly lower variance, resulting in more consistent completion times among students and reducing the likelihood of faster students waiting for slower ones.

e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future?

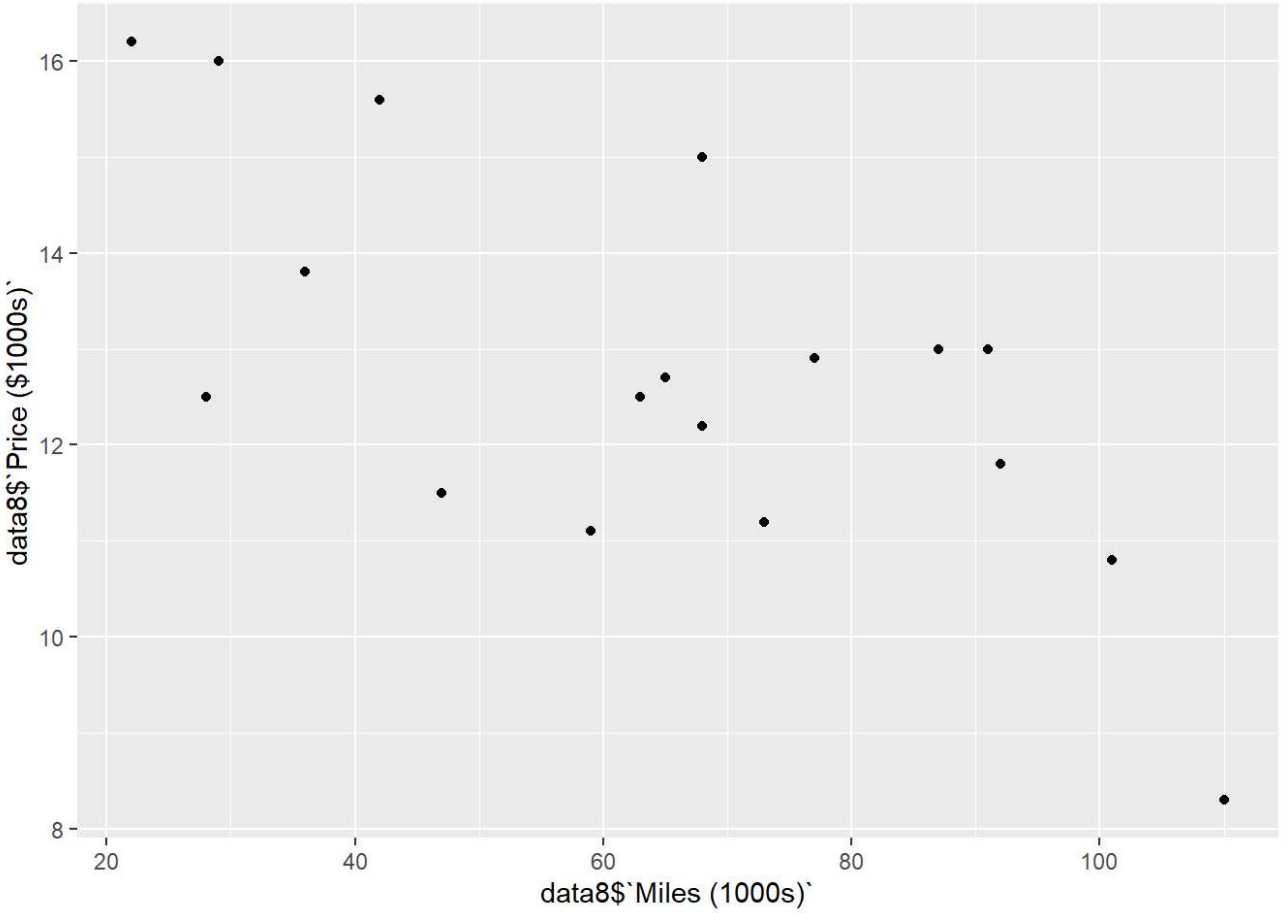
   An end-of-program exam for both groups could determine if the learning outcomes differ between the methods.

Question #8

The Toyota Camry is one of the best-selling cars in North America. The cost of a previously owned Camry depends upon many factors, including the model year, mileage, and condition. To investigate the relationship between the car's mileage and the sales price for a 2007 model year Camry, Attached data file Camry show the mileage and sale price for 19 sales (Pricehub website, February 24, 2012).

a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the vertical axis.

```
ggplot(data8)+
  geom_point(aes(data8$`Miles (1000s)`,data8$`Price ($1000s)`))
```



b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

The two variables seem negatively related, approximating a straight line.

c. Develop the estimated regression equation that could be used to predict the price ($1000s) given the miles (1000s).

The estimated regression equation is Price = 16.4698 - 0.0588 * miles.

```
lm_camry <- lm(data8$`Price ($1000s)` ~ data8$`Miles (1000s)`, data = data8)

summary(lm_camry)
```

```
##
## Call:
## lm(formula = data8$`Price ($1000s)` ~ data8$`Miles (1000s)`,
##     data = data8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32408 -1.34194  0.05055  1.12898  2.52687
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           16.46976    0.94876  17.359 2.99e-12 ***
## data8$`Miles (1000s)` -0.05877    0.01319  -4.455 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.541 on 17 degrees of freedom
## Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
## F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

d. Test for a significant relationship at the .05 level of significance.

   The p-value is 0.000348, which is smaller than the level of significance of 0.05.

e. Did the estimated regression equation provide a good fit? Explain.

   The multiple R-squared is 0.539, which indicates that it's a good fit.

f. Provide an interpretation for the slope of the estimated regression equation.

   The regression slope is -.0588, meaning a one-unit x increase leads to a .0588 y decrease. In the context of the problem, The regression slope means that each extra 1000 miles reduces the predicted car price by $58.8.

g. Suppose that you are considering purchasing a previously owned 2007 Camry that has been driven 60,000 miles. Using the estimated regression equation developed in part (c), predict the price for this car. Is this the price you would offer the seller.

```
16.4698 -.0588*60
```

```
## [1] 12.9418
```

   The predicted price for a 2007 Camry with 60,000 miles is $12.9418, and it might be a good suggestion offerring the seller considering other factors like cars conditions.

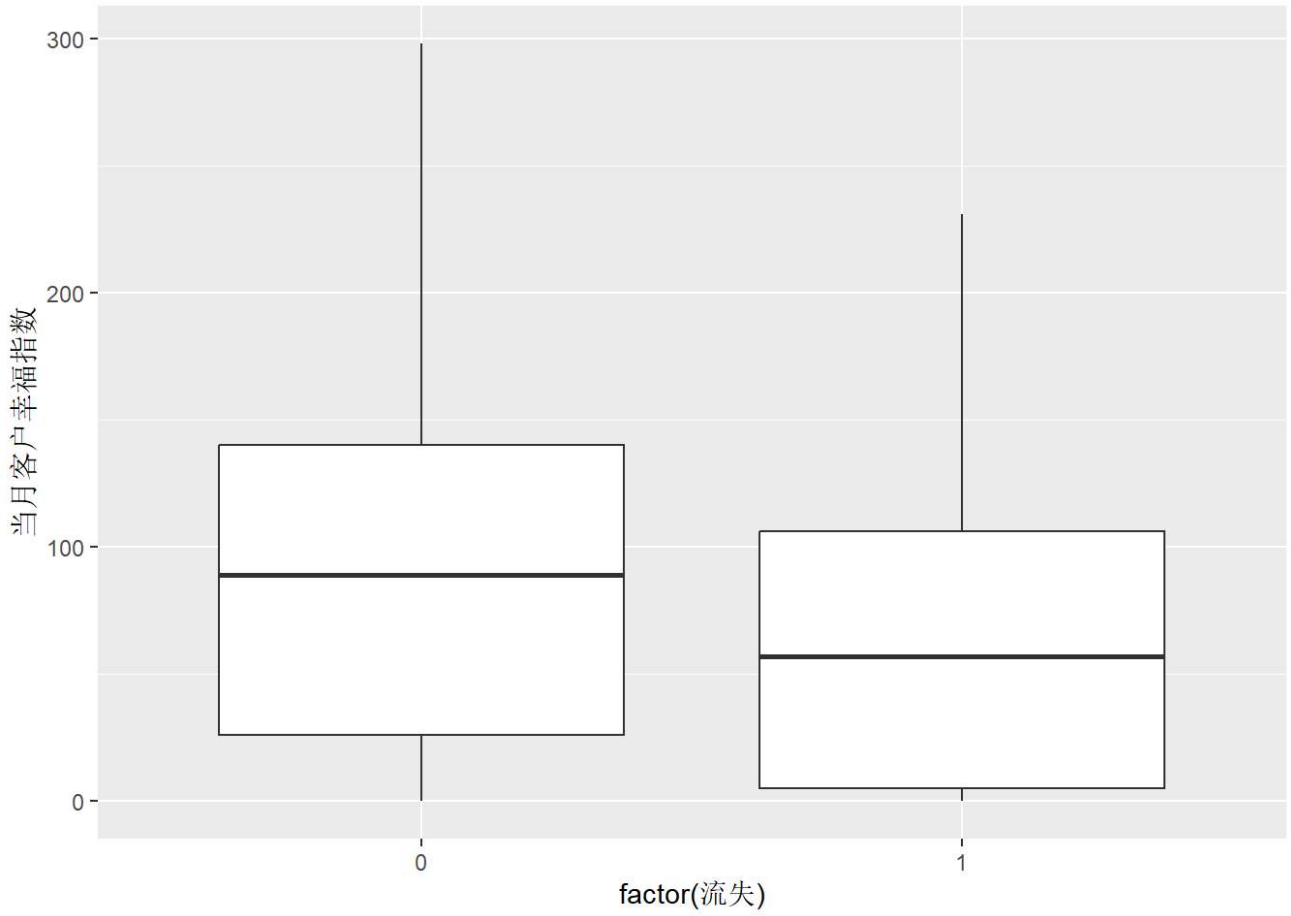
Question #9

附件WE.xlsx是某提供网站服务的Internet服务商的客户数据。数据包含了6347名客户在11个指标上的表现。其中"流失"指标中0表示流失，"1"表示不流失，其他指标含义看变量命名。

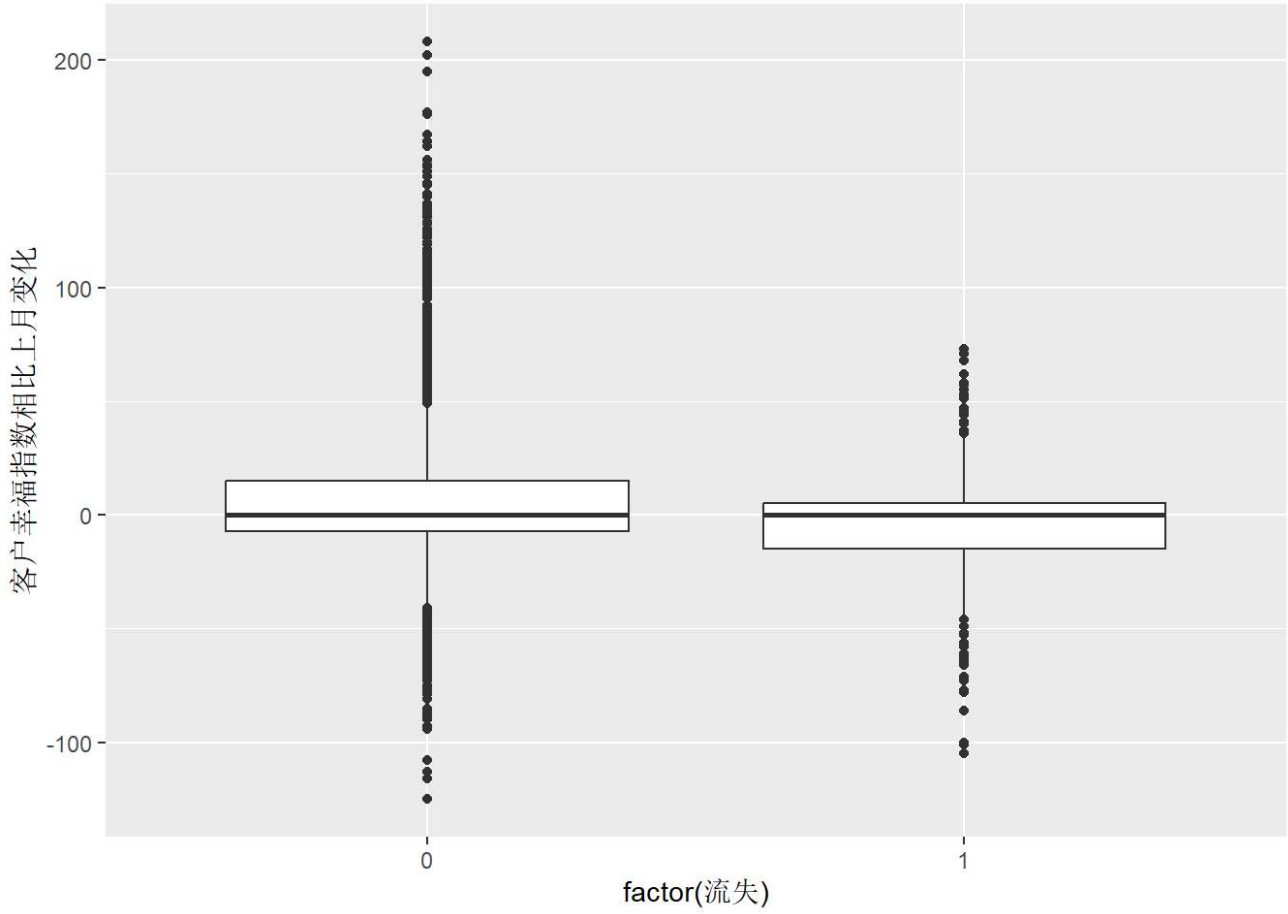
a. 通过可视化探索流失客户与非流失客户的行为特点（或特点对比），你能发现流失与非流失客户行为在哪些指标有可能存在显著不同？

   对是否流失客户与其他特征绘制箱型图，观察可知各变量中当月客户幸福指数、客户幸福指数相比上月变化、当月服务优先级、客户使用期限、访问间隔变化可能存在显著不同，需要进一步计算。

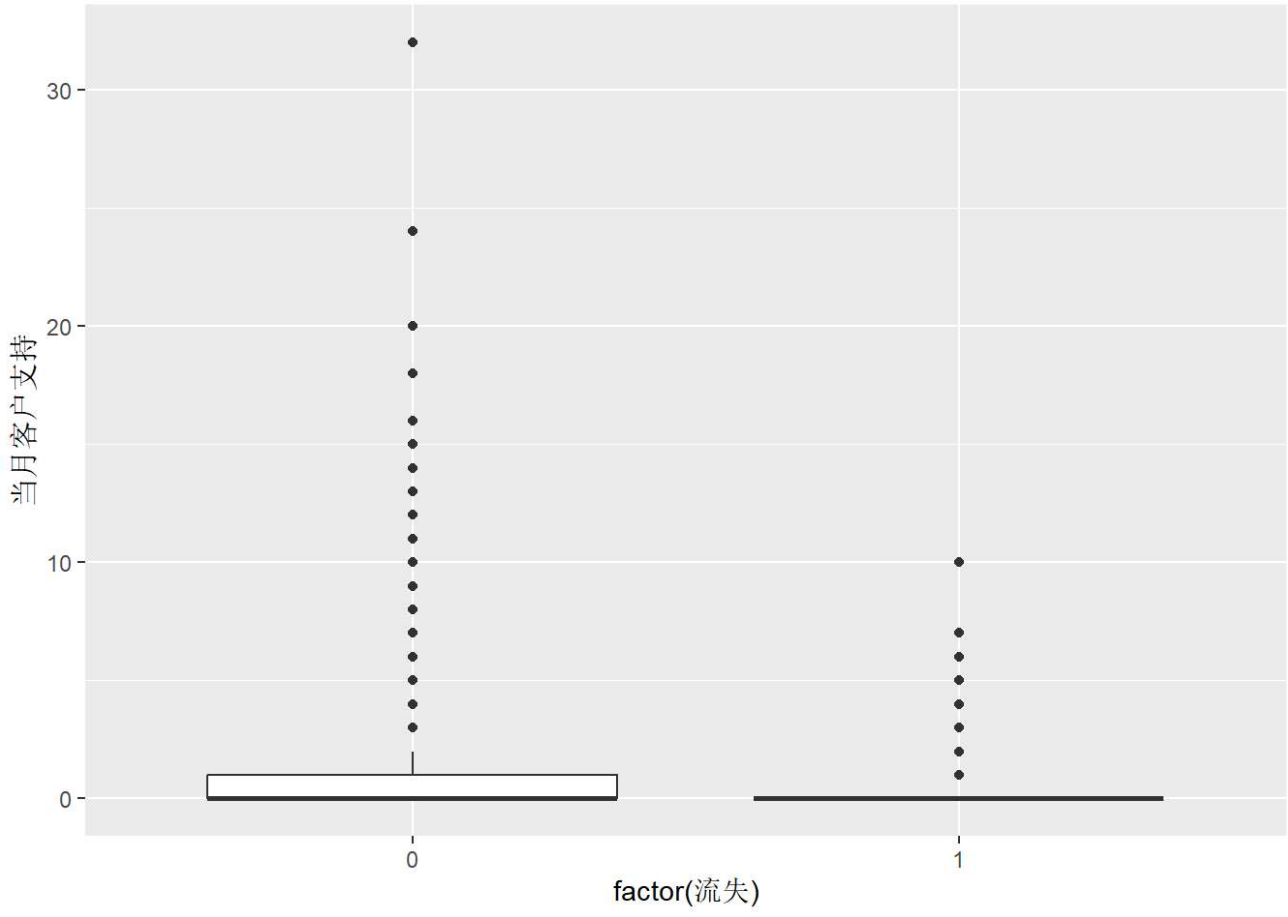
```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失), y=当月客户幸福指数))
```



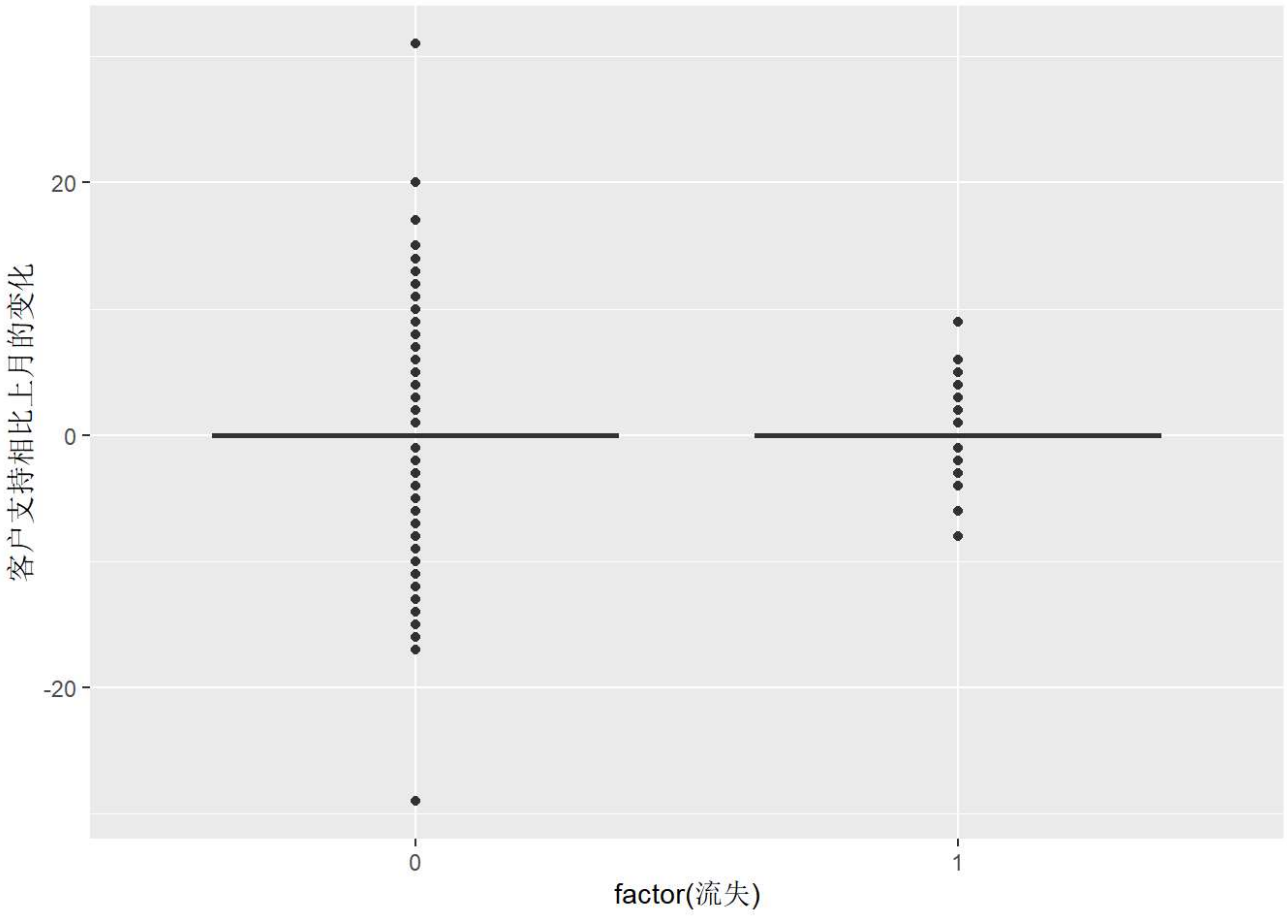

```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=客户幸福指数相比上月变化))
```



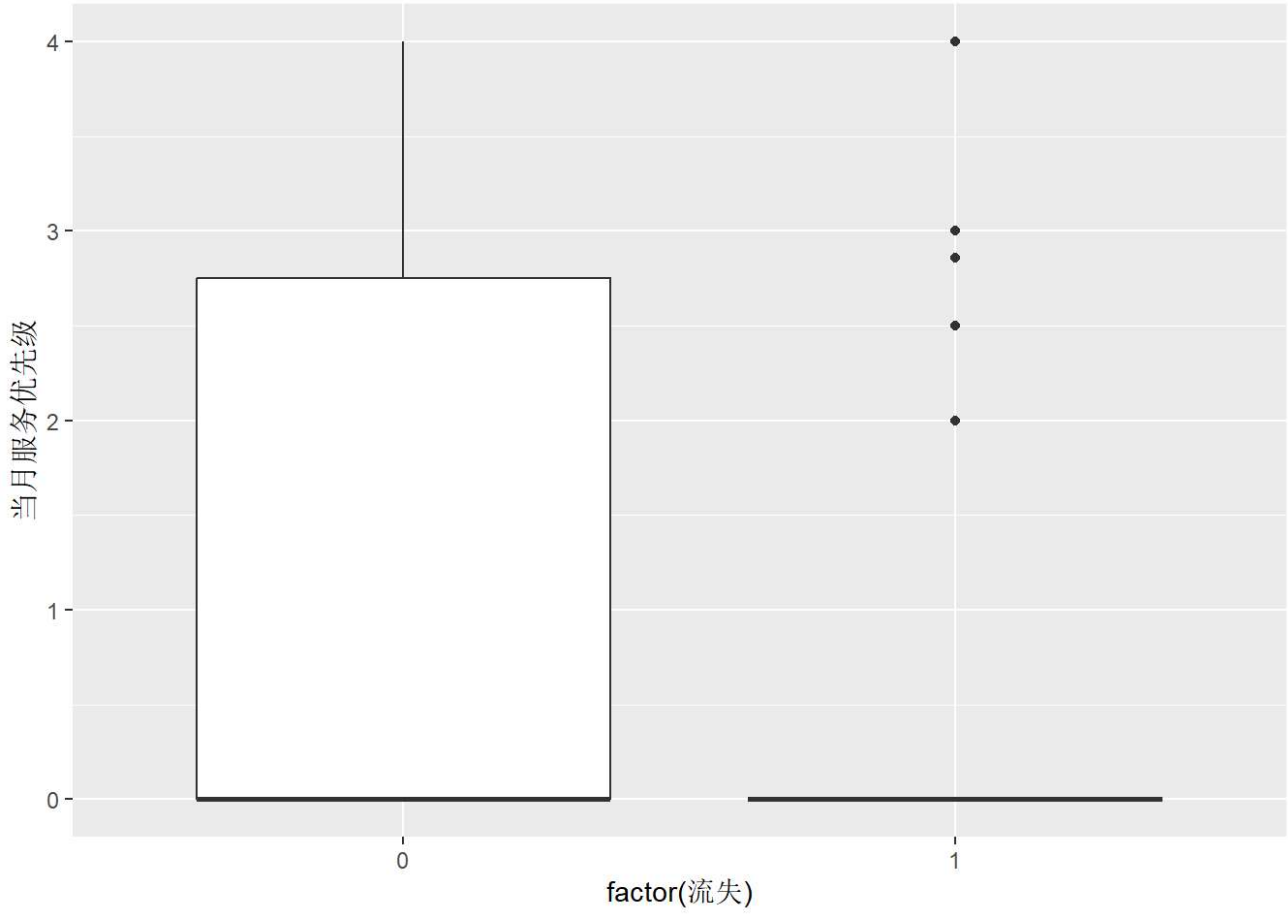

```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=当月客户支持))
```



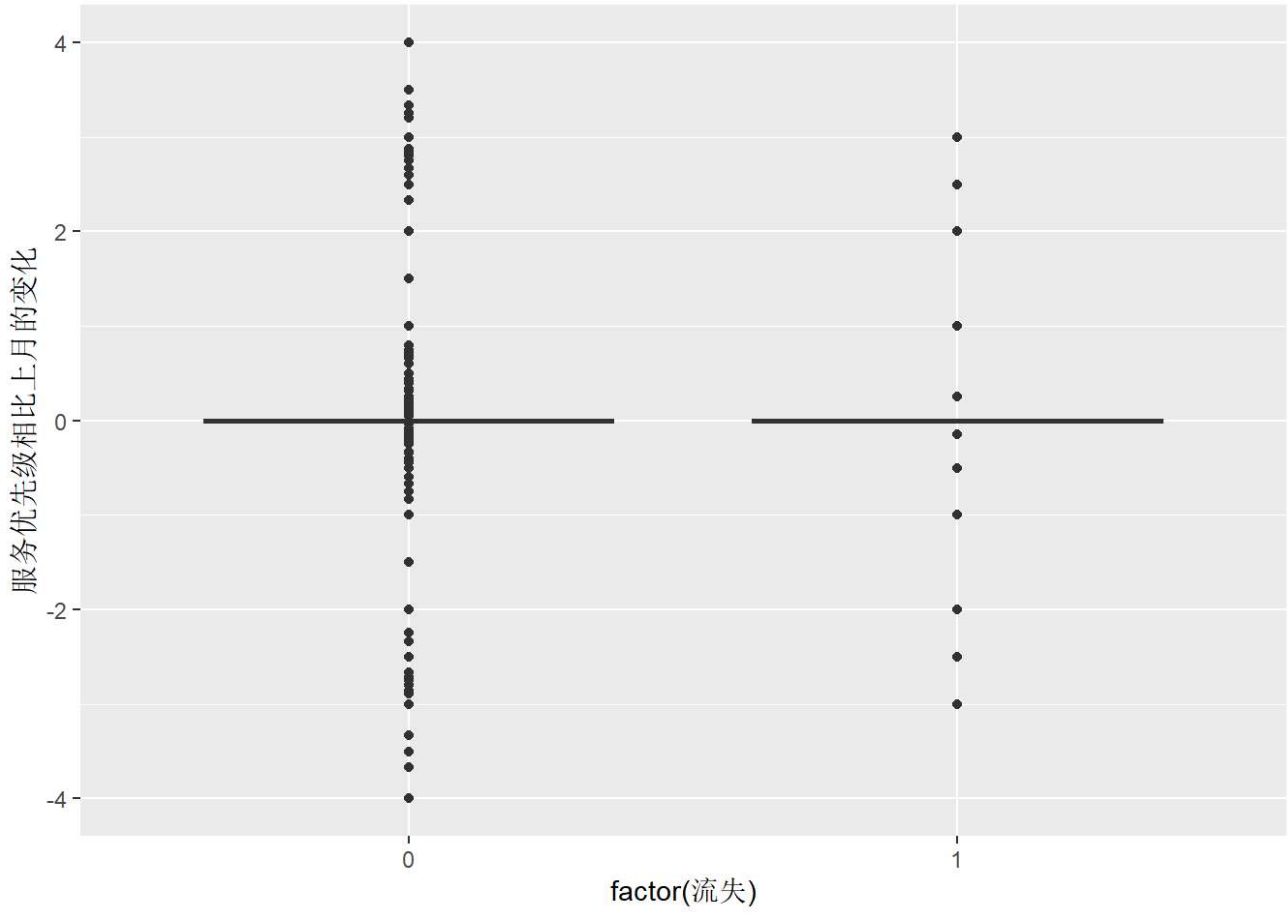
```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=客户支持相比上月的变化))
```



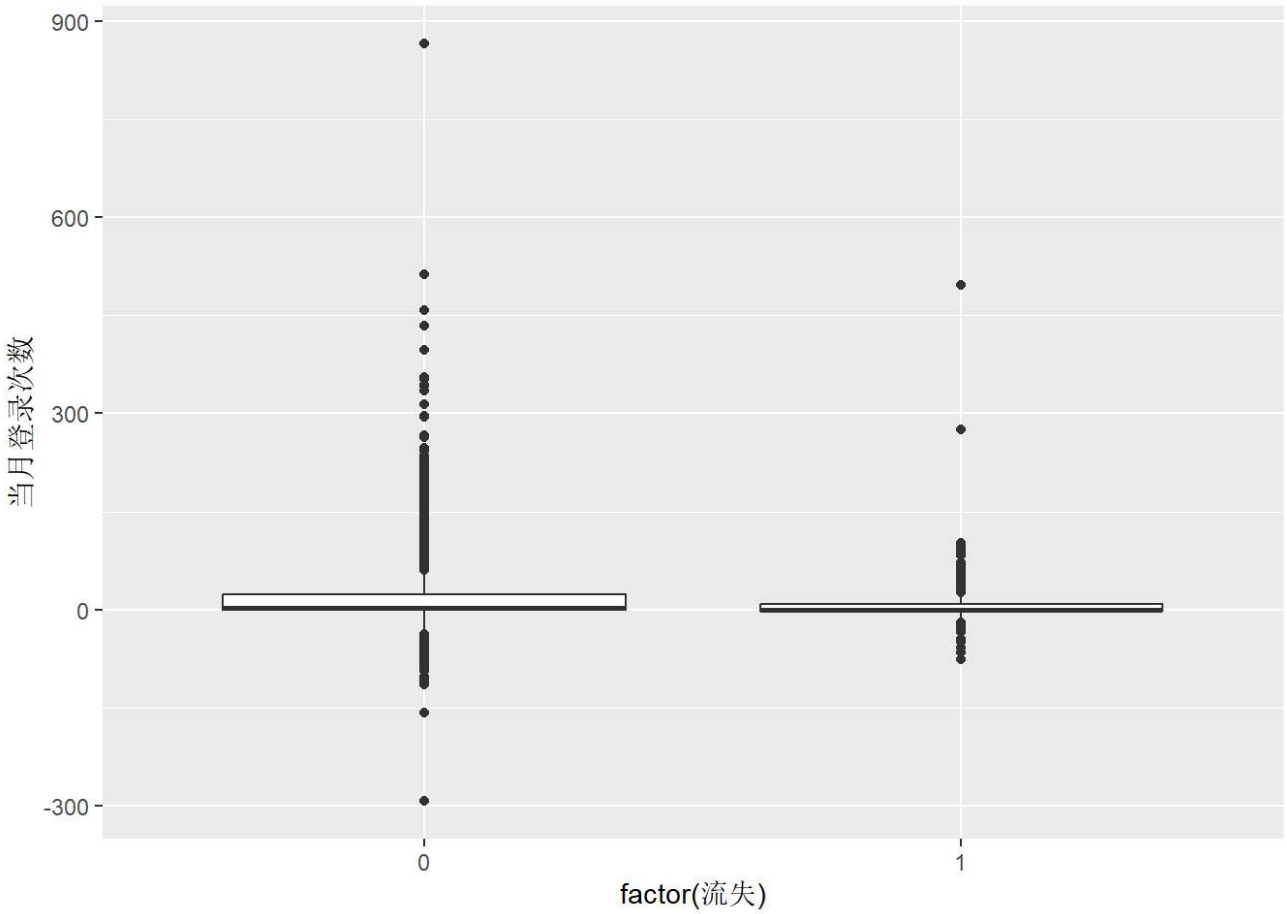

```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=当月服务优先级))
```



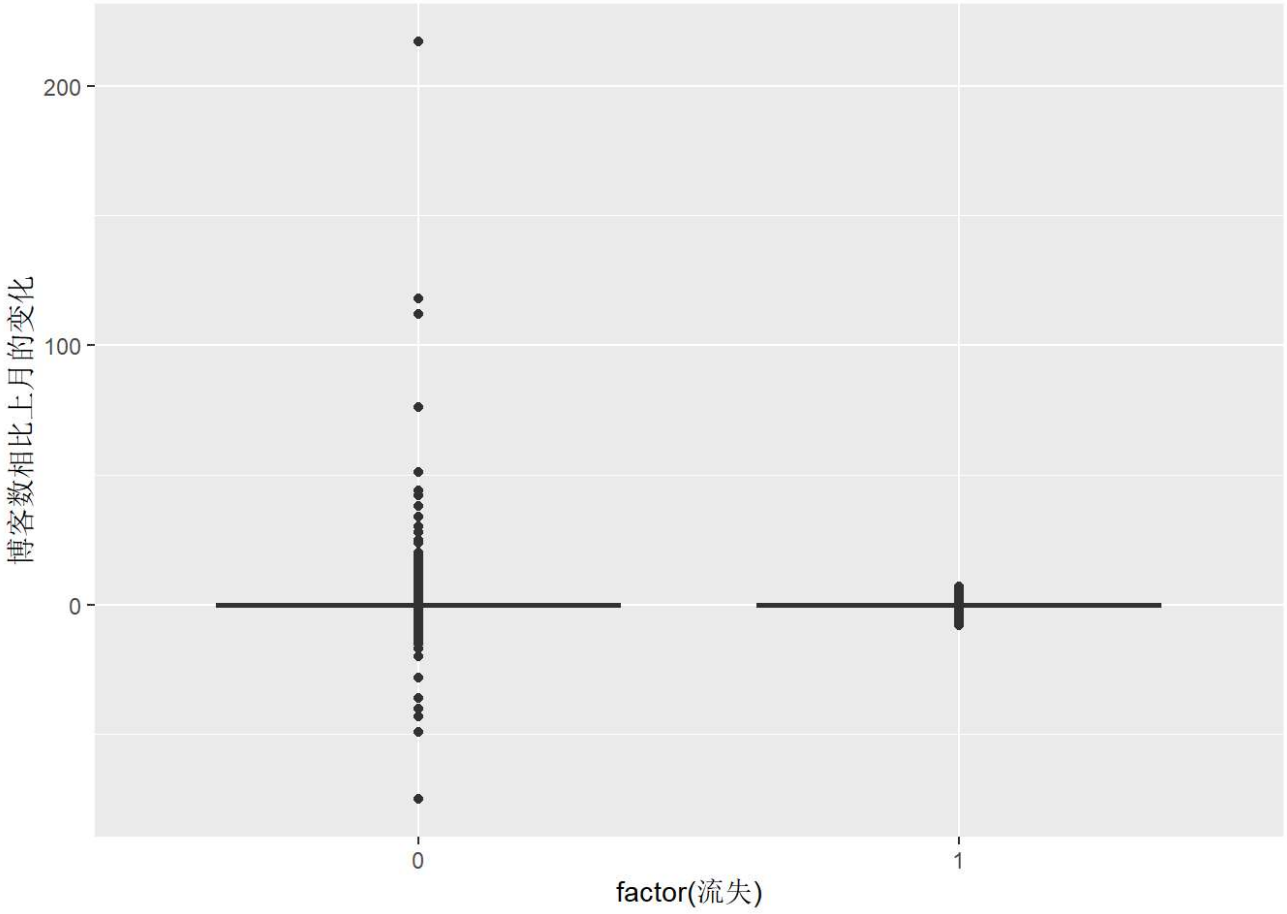

```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=服务优先级相比上月的变化))
```



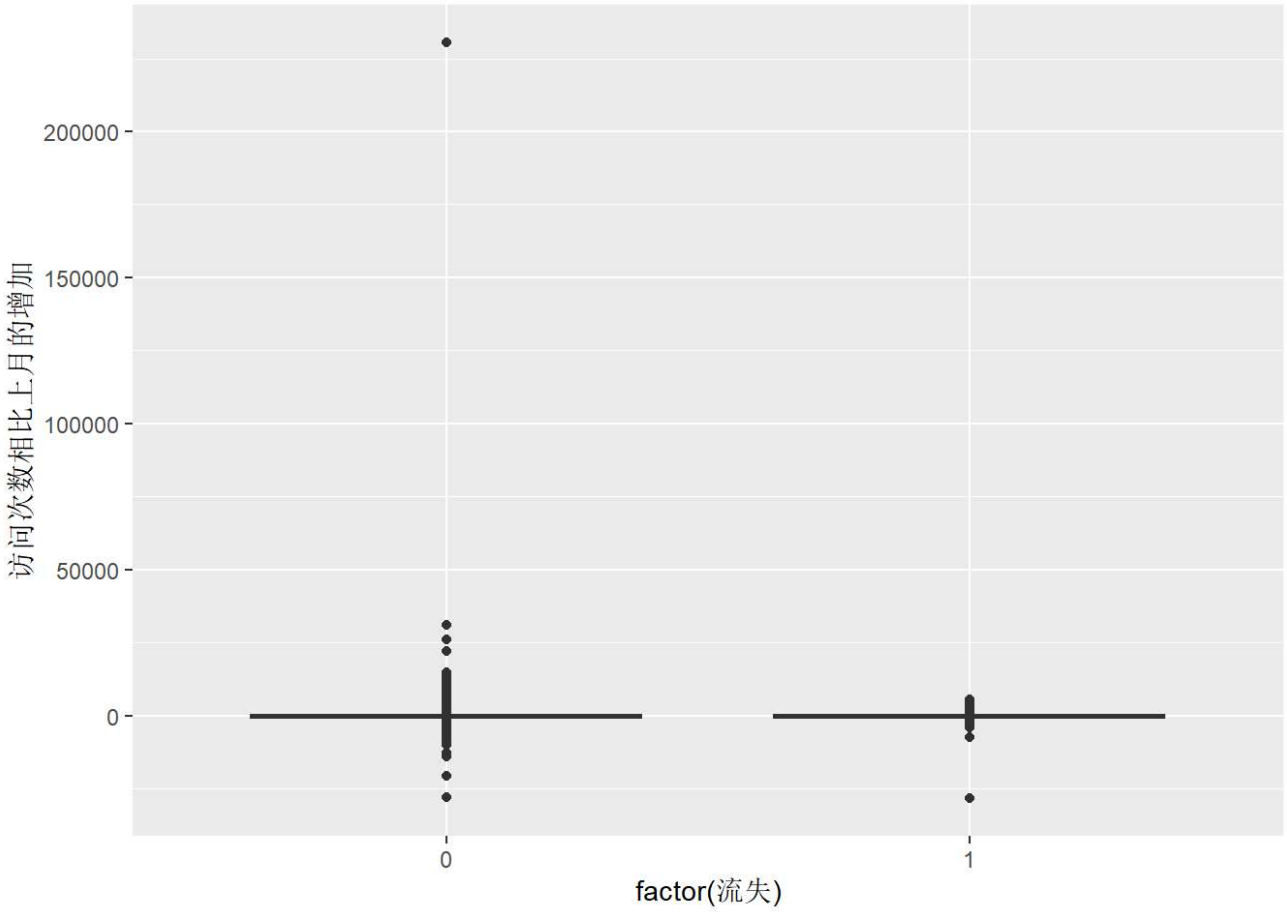
```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=当月登录次数))
```



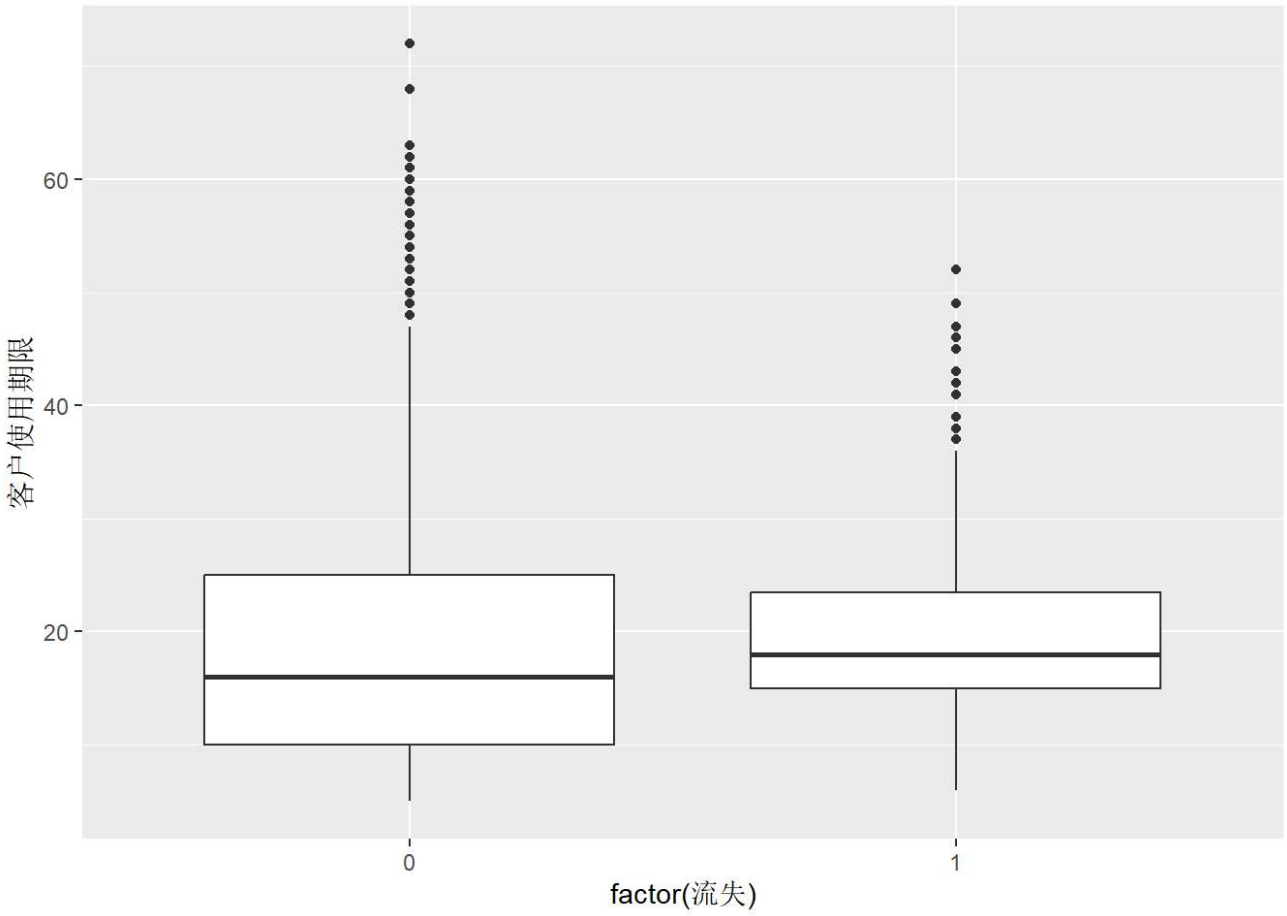

```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=博客数相比上月的变化))
```



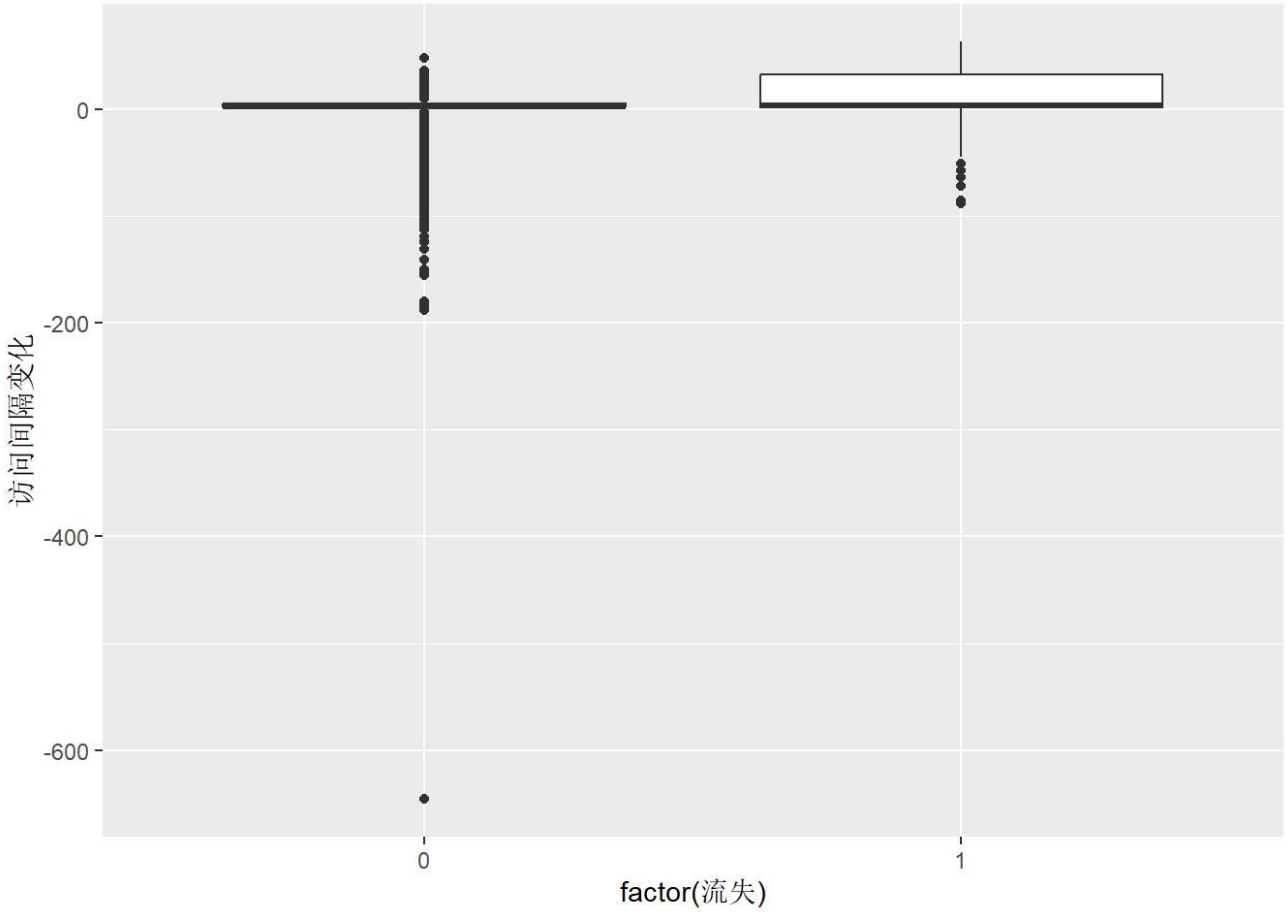

```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=访问次数相比上月的增加))
```



```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=客户使用期限))
```




```
ggplot(data9)+
  geom_boxplot(aes(x=factor(流失),y=访问间隔变化))
```



b. 通过均值比较的方式验证上述不同是否显著。

以0.05为显著性水平，t.test的结果可知

显著的指标有：当月客户幸福指数、客户幸福指数相比上月变化、当月客户支持、当月服务优先级、当月登录次数、博客数相比上月的变化、客户使用期限、访问间隔变化

不显著的指标有：客户支持相比上月的变化、服务优先级相比上月的变化、访问次数相比上月

```
t.test(data9$流失,data9$当月客户幸福指数)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$当月客户幸福指数
## t = -104.89, df = 6346.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -88.89678 -85.63481
## sample estimates:
##   mean of x   mean of y
##  0.05089018 87.31668505
```

```
t.test(data9$流失,data9$客户幸福指数相比上月变化)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$客户幸福指数相比上月变化
## t = -12.941, df = 6346.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.766322 -4.249118
## sample estimates:
##  mean of x  mean of y
## 0.05089018 5.05861037
```

```
t.test(data9$流失,data9$当月客户支持)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$当月客户支持
## t = -30.046, df = 6552.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6981912 -0.6126643
## sample estimates:
##  mean of x  mean of y
## 0.05089018 0.70631795
```

```
t.test(data9$流失,data9$客户支持相比上月的变化)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$客户支持相比上月的变化
## t = 2.4454, df = 6521.1, p-value = 0.0145
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01146925 0.10417594
## sample estimates:
##    mean of x    mean of y
##  0.050890184 -0.006932409
```

```
t.test(data9$流失,data9$当月服务优先级)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$当月服务优先级
## t = -45.341, df = 6697.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7948311 -0.7289510
## sample estimates:
##  mean of x  mean of y
## 0.05089018 0.81278121
```

```
t.test(data9$流失,data9$服务优先级相比上月的变化)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$服务优先级相比上月的变化
## t = 1.1178, df = 6633.4, p-value = 0.2637
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0156167  0.0570591
## sample estimates:
##  mean of x  mean of y
## 0.05089018 0.03016898
```

```
t.test(data9$流失,data9$当月登录次数)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$当月登录次数
## t = -29.653, df = 6346.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.71342 -14.64060
## sample estimates:
##   mean of x   mean of y
##  0.05089018 15.72790295
```

```
t.test(data9$流失,data9$博客数相比上月的变化)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$博客数相比上月的变化
## t = -1.8159, df = 6374.2, p-value = 0.06943
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.22115719  0.00845828
## sample estimates:
##  mean of x  mean of y
## 0.05089018 0.15723964
```

```
t.test(data9$流失,data9$访问次数相比上月的增加)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$访问次数相比上月的增加
## t = -2.4327, df = 6346, p-value = 0.01501
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -173.8289  -18.6904
## sample estimates:
##   mean of x   mean of y
##  0.05089018 96.31054041
```

```
t.test(data9$流失,data9$客户使用期限)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$客户使用期限
## t = -134.51, df = 6350.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.12057 -18.57125
## sample estimates:
##   mean of x   mean of y
##  0.05089018 18.89680164
```

```
t.test(data9$流失,data9$访问间隔变化)
```

```
##
##  Welch Two Sample t-test
##
## data:  data9$流失 and data9$访问间隔变化
## t = -16.467, df = 6347.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.155834 -3.271612
## sample estimates:
##  mean of x  mean of y
## 0.05089018 3.76461320
```

c. 以"流失"为因变量，其他你认为重要的变量为自变量（提示：a、b两步的发现），建立回归方程对是否流失进行预测。

```
lm_we <- glm(流失 ~ 当月客户幸福指数+客户幸福指数相比上月变化+当月客户支持+当月服务优先级+当月登录次数+博客数相比上月的变化+客户使用期限+访问间隔变化, data = data9, family = binomial)
summary(lm_we)
```

```
##
## Call:
## glm(formula = 流失 ~ 当月客户幸福指数 + 客户幸福指数相比上月变化 +
##     当月客户支持 + 当月服务优先级 + 当月登录次数 +
##     博客数相比上月的变化 + 客户使用期限 + 访问间隔变化,
##     family = binomial, data = data9)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.8763327  0.1212590 -23.721  < 2e-16 ***
## 当月客户幸福指数         -0.0051988  0.0011558  -4.498 6.86e-06 ***
## 客户幸福指数相比上月变化 -0.0093063  0.0024124  -3.858 0.000114 ***
## 当月客户支持             -0.0221691  0.0714550  -0.310 0.756369
## 当月服务优先级           -0.0447524  0.0741355  -0.604 0.546072
## 当月登录次数              0.0008545  0.0019376   0.441 0.659211
## 博客数相比上月的变化     -0.0009717  0.0205099  -0.047 0.962213
## 客户使用期限              0.0142559  0.0052396   2.721 0.006513 **
## 访问间隔变化              0.0169505  0.0042787   3.962 7.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2452.2  on 6338  degrees of freedom
## AIC: 2470.2
##
## Number of Fisher Scoring iterations: 6
```

d. 根据上一步预测的结果，对尚未流失（流失=0）的客户进行流失可能性排序，并给出流失可能性最大的前100名用户ID列表。

```
data9 %>%
  add_predictions(lm_we,type = "response") %>%
  arrange(desc(pred)) %>%
  filter(流失 == 0) %>%
  slice_head(n=100)
```

```
## # A tibble: 100 × 14
##    客户ID  流失 当月客户幸福指数 客户幸福指数相比上月变化 当月客户支持
##     <dbl> <dbl>            <dbl>                  <dbl>        <dbl>
##  1   109     0                0                   -125            0
##  2  1971     0                0                   -113            0
##  3     1     0                0                      0            0
##  4  2076     0               29                    -69            0
##  5    14     0                0                      0            0
##  6    76     0                1                    -70            0
##  7     3     0                0                      0            0
##  8    18     0                0                      0            0
##  9    21     0                0                      0            0
## 10  2244     0               16                    -38            0
## # ℹ 90 more rows
## # ℹ 9 more variables: 客户支持相比上月的变化 <dbl>, 当月服务优先级 <dbl>,
## #   服务优先级相比上月的变化 <dbl>, 当月登录次数 <dbl>,
## #   博客数相比上月的变化 <dbl>, 访问次数相比上月的增加 <dbl>,
## #   客户使用期限 <dbl>, 访问间隔变化 <dbl>, pred <dbl>
```